Wide-Area Performance Profiling of 10GigE and InfiniBand Technologies

Nageswara S. V. Rao, Weikuan Yu, William R. Wing, Stephen W. Poole, and Jeffrey S. Vetter
Oak Ridge National Laboratory, Oak Ridge, TN 37831
{raons,wyu,wrw,spoole,vetter}@ornl.gov

***Abstract*—For wide-area high-performance applications, light-paths provide 10Gbps connectivity, and multi-core hosts with PCI-Express can drive such data rates. However, sustaining such end-to-end application throughputs across connections of thousands of miles remains challenging, and the current performance studies of such solutions are very limited. We present an experimental study of two solutions to achieve such throughputs based on: (a) 10Gbps Ethernet with TCP/IP transport protocols, and (b) InfiniBand and its wide-area extensions. For both, we generate performance profiles over 10Gbps connections of lengths up to 8600 miles, and discuss the components, complexity, and limitations of sustaining such throughputs, using different connections and host configurations. Our results indicate that IB solution is better suited for applications with a single large flow, and 10GigE solution is better for those with multiple competing flows.**

Keywords and Phrases: InfiniBand, 10GigE LAN/WAN-PHY, TCP performance, high-performance networking.

I. Introduction

A number of large-science and data-center applications require efficient data transport with throughput rates in the range of 10Gbps over wide-area network connections. For example, the transfer of a terabyte dataset at a sustained rate of 9Gbps takes about 15 minutes, whereas a petabyte dataset takes about 10 days. However, data transfers with such end-to-end performance are still not being widely realized over wide-area production networks, despite the availability of both wide-area connections and hosts capable of such rates. To support such applications, dedicated 10Gbps connections can be provisioned over several networks, including Internet2 [1], ESnet's Data Sciences Network [2], UltraScience Net [3], CHEETAH [4], and others. At the same time, multi-core hosts equipped with PCI-Express I/O bus achieve 9.9 Gbps throughput for local-area data transfers using 10-Gigabit Ethernet (10GigE) Network Interface Cards (NIC). Also, there has been an active development of high performance protocols such as Binary Increase Congestion Control (BIC) [5], CUBIC [6], High-Speed TCP (HSTCP) [7], Hamilton TCP (HTCP) [8], Scalable TCP [9] and others [10], [9], [11] that target such data rates over wide-area connections. These solutions are often seen in higher-level transport tools such as bbcp [12] and GridFTP [13]. While the ingredients for 10Gbps throughput are available, the task of sustaining end-to-end throughput at such rates over thousands of miles still remains complex, and the performance measurements on real connections that pinpoint the pertinent technical issues are rather limited.

The data transport across wide-area networks has traditionally been based on 1/10GigE technologies combined with SONET or WAN-PHY technologies in the wide-area. InfiniBand (IB) was originally developed for data transport over enterprise-level interconnections for clusters, supercomputers and storage systems. It is quite common to achieve data transfer rates of 7.5 Gbps using commodity IB Host Channel Adapters (HCA) (SDR 4X, 8 Gbps peak) by simply connecting them to IB switches. However, geographically separated IB deployments still rely on transition to TCP/IP and its ability to sustain 7.0-8.0 Gbps rates for wide-area data transfers, which by itself requires significant per-connection optimization. Very recently, there have been hardware implementations of InfiniBand over Wide-Area (IBoWA) devices, in particular Longbow XR from Obsidian Research

SC2008 November 2008, Austin, Texas, USA 978-1-4244-2835-9/08 $25.00 ©2008 IEEE

Corporation [14] and NX5010ae from Network Equipment Technologies [15]. Preliminary results indicate that IBoWA technologies can sustain multiple Gbps transfer rates over thousands of miles, for example, 7 Gbps over 8600-mile connections [16], [17]. These devices can be simply dropped-in at the edges of wide-area connections, thereby by-passing the transition to TCP/IP, and somewhat surprisingly offer a potential alternate solution for wide-area data transport.

There have been very limited number of results published on the wide-area measurements of 10GigE and IB technologies for the problem space described above, and much less seems to be known about their comparative performance. While both these technologies can be deployed on SONET or 10GigE wide-area connections, their deployment costs, configurations and regions of optimal performance are significantly different. Thus, it is critical to correlate and understand the performances of these two competing solutions to make informed choices, since once deployed, significant costs are involved in replacing one by the other.

We conduct structured experiments to assess the throughput performance of these two technologies in terms of their scalability to connection lengths and stability to repeated measurements. We consider various combinations of hosts, edge devices and connection modalities and lengths. We allocate approximately a few days of effort in tuning the configuration for each of these methods, and utilize openly available software implementations. Our choice is mainly motivated by the likely performance achieved by an informed user with limited efforts rather than the peak performance achievable by domain experts with considerably more efforts (weeks to months) [1]. In comparison, these two methods are suitable to different scenarios, and represent different cost-benefit trade-offs. The 10GigE solution requires a transport method such as TCP to ensure reliable data transport. This solution is generally preferred if there are several data flows each with multiple Gbps flow rate, since congestion control of TCP provides a graceful scale-down of individual flows. On the other hand, significant effort is needed in choosing and tuning an appropriate TCP method to sustain multiple Gbps flows at thousands of miles. As our measurements indicate a good solution requires a suitable number of streams optimized for each connection length.

IBoWA solution is best suited to scenarios with a single large multiple Gbps flow with the rest of IP traffic aggregated to a single 1 Gbps flow. However, it would require dedicated wide-area connections particularly when SONET is used. Limited amounts of wide-area cross-traffic may be supported over WAN-PHY connections, but high levels lead to lower performance, and thus this solution does not provide a graceful performance scale-down. On the other hand, other traffic at an aggregate rate of 1 Gbps can be supported by using the additional ports provided by IBoWA devices without affecting their IB performance. Furthermore, IBoWA is particularly attractive for its capability to natively extend the IB interconnects of supercomputers and clusters, and to support certain MPI applications over wide-area.

This paper is organized as follows. We first describe our experimental environments in Section II. We describe the 10GigE and IBoWA solutions in Sections III and IV, respectively. We compare the performance of these two methods in Section V, and present our conclusions in VI.

II. Test Environments

A. Host and Edge Systems

We consider three types of host configurations shown in Table I. We primarily utilize pairs of quad-core dual-socket and dual-core dual-socket hosts with 2.0 and 2.6 GHz processors, respectively, and each is equipped with 8x PCI-Express I/O bus. These hosts run Linux 2.6.18 and 2.6.23 kernels, respectively, both of which support auto-tuning and pluggable TCP congestion control modules. These hosts are equipped with Myrinet 10GigE cards, and InfiniBand HCAs from Voltaire or Mellanox. For the sake of comparison with less powerful hosts, we also utilize a pair of dual-core single-socket 2.19 GHz Xeon processor hosts with PCI-X bus with 2.6.9 Linux

[1]The incremental performance improvements of these methods possible by further optimizations may not be "simply" extrapolated from our results here; in particular, much higher TCP performance may be possible by expert-level optimizations, which need to be performed on per-connection basis with significantly more effort.

TABLE I
HOST CONFIGURATIONS

	hosts	processors	10GigE NIC	PCI bus	IB HCA	Linux
I	quad-core dual-socket	2.0 GHz Opteron	Myrinet	PCI-Express 400EX	Voltaire	2.6.16
II	dual-core dual-socket	2.6 GHz Xeon	Myrinet	PCI-Express	Mellanox 4X 4X DDR Connect X	2.6.23
III	dual-core single-socket	2.19 GHz Xeon	Netrion 1	PCI-X 2.0	n/a	2.6.9

kernel, which does not fully support TCP auto-tuning capability. These hosts are equipped with Netrion NICs connected via PCI-X 2.0 bus.

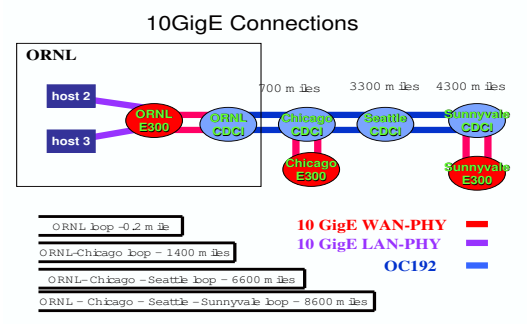
At the edges of wide-area connections, we utilize Longbow XR units to support IB over SONET OC192, and also over 10GigE WAN-PHY and LAN-PHY connections. In a general configuration, IB ports on Longbow XR devices are connected to Flextronics and Cisco IB switches, which are locally connected to two separate IB clusters with their own IB Subnet Managers (SM). We have also conducted experiments using two NX5010ae devices for IB extension over SONET, where the Cisco IB switch running its internal SM is connected to one NX5010ae and the other is directly connected to a host. Results are available in an earlier paper [17]. For even simpler connections, hosts are directly connected to IB ports on IBoWA devices with SM running on one of them.

B. Wide-Area Connections

The wide-area connections are dynamically configured as SONET or WAN-PHY on UltraScience Net [16]. USN data-plane consists of four thousand miles of dual OC192 (9.6Gbps) connections spanning Oak Ridge, Chicago, Seattle and Sunnyvale as shown in Figure 1(a). SONET links are provided between Ciena CDCI switches, which in turn are connected to 10GigE WAN-PHY ports of Force10 E300 Ethernet switches as shown in Figure 1(b). The E300 10GigE ports can be configured as WAN-PHY or LAN-PHY, and the two can be cross-connected to provide the conversion between the two.

We utilize SONET OC192 and 10GigE WAN-PHY connections of lengths 0.2, 1400, 6600 and 8600 miles with RTTs (Round Trip Time) shown in Table II. Hosts

(a) UltraScience Net



(b) 10Gbps switching at CDCI and E300

Fig. 1. Provisioning of 10Gbps connections on UltraScience Net.

and IBoWA devices are installed at Oak Ridge National Laboratory (ORNL) in pairs to dynamically provision loop-back connections of various lengths and types. The 10GigE NICs on hosts are connected to LAN-PHY ports on E300. Longbow XRs are connected to CDCIs for SONET connections, and to E300 for 10GigE connections and are switched in software between WAN-PHY and LAN-PHY. For the wide-area connections, SONET loop-back connections are realized by employing pairs

TABLE II
RTT OF DIFFERENT CONNECTION LENGTHS

connection length (miles)	0.2	1400	6600	8600
RTT (ms)	0.28	26.8	128	163

of OC192 links switched exclusively at CDCIs. Similarly, WAN-PHY or LAN-PHY connections between two E300 ports are realized by switching at E300s. For 1400-mile WAN-PHY connection, two parallel WAN-PHY connections are provisioned between ORNL and Chicago by terminating the OC192 links between CDCI switches at the corresponding E300 switches; then the two connections are cross-connected on Chicago E300 switch. The longest 8600-mile WAN-PHY loop-back connection between two ports of ORNL E300 is provisioned as follows: (i) ORNL to Chicago OC192 link is terminated on E300s at both ends on their 10GigE ports, (ii) Chicago to Sunnyvale OC192 connection through Seattle CDCI is similarly terminated on E300s at both ends, and (iii) a parallel ORNL to Sunnyvale OC192 connection through Chicago and Seattle CDCIs is terminated on respective E300 10GigE ports. Then the connections (i) and (ii) are cross-connected on Chicago E300, and connections (ii) and (iii) are cross-connected on Sunnyvale E300.

III. TCP OVER 10GIGE SOLUTIONS

The 10GigE data transport solution consists of hosts connected to LAN-PHY ports on ORNL E300, which are in turn cross-connected to wide-area WAN-PHY connections described in the previous section. The throughput measurements are collected using iperf tool. Using Linux sysctl utilities, we dynamically selected different TCP congestion control protocols, including the default BIC, HTCP, HSTCP, Scalable TCP, and Vegas. To keep the presentation tractable, we present measurement results for the top two protocols in terms of throughput, namely BIC and HTCP; the others performed significantly lower even at 1400 miles, as shown in Figure 4(a) for the top three protocols. In all cases, TCP auto-tuning was turned on, and our limited attempts to manually tune these TCP parameters did not improve the performance significantly.

A. Performance Profiles

Let $T_A(d,n)$, $A \in \{BIC, HTCP\}$ denote the iperf throughput measurement over 10GigE WAN-PHY connection of length d using n TCP streams of type A. Let $\bar{T}_A(d)$ denote the average throughput over 10 repeated measurements using TCP of type A over a connection of length d with an appropriately chosen value for n or averaged over a chosen range of values for n. We compute the *Decrease Per Mile* (DPM) of the throughput for connection of length d_i miles with respect to the connection of same type of base length d_0 miles as $D_A(d_i) = \frac{\bar{T}_A(d_0) - \bar{T}_A(d_i)}{d_i - d_0}$.

We characterize the throughput performance of 10GigE solution based on the following profiles:

(a) **Distance Scalability:** The *distance-profile* is generated by measuring TCP throughputs $T_A(d,n)$ for $A \in \{BIC, HTCP\}$, $n = 1, 2, \ldots, 20$ and $d = 0.2, 1400, 6600, 8600$ miles. The distance-profile can be illustrated as a two-dimensional surface $T_A(d,n)$ with d and n along x and y axes, respectively.

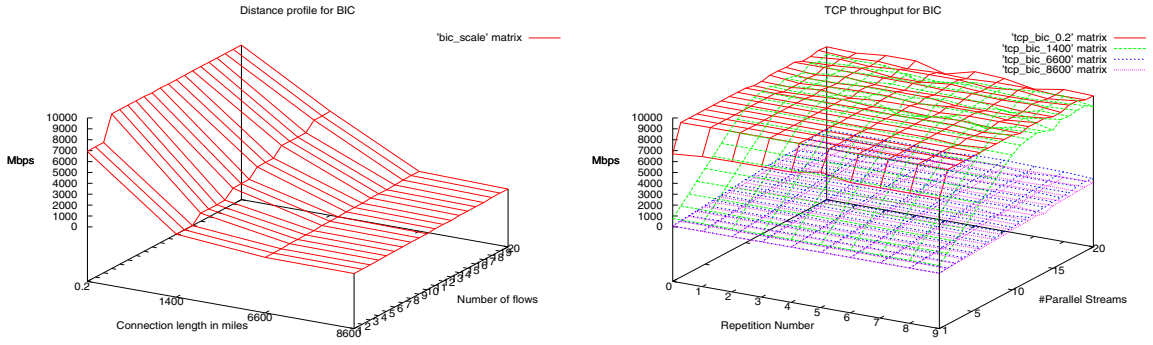
(b) **Performance Stability:** The *stability-profile* is generated by repeating the throughput measurements $T_A(d,n)$ ten times for fixed d and $n = 1, 2, \ldots, 20$. The stability-profile is also illustrated as a two-dimensional surface, where x axis represents the repetition number and y axis represents n. We also compute $\bar{T}_A(d)$ averaged over the repeated measurements and $n = 5, 6, \ldots, 15$, and then compute the corresponding DPM $D_A(d)$ values for base length $d_0 = 0.2$ and $d = 1400, 6600, 8600$ miles.

The two-dimensional plots of these profiles visually depict the overall qualitative behavior of the throughput as we vary the connection length and the number of parallel streams, and repeat the experiments.

B. Transport Methods

TCP protocols targeting connections of thousands of miles with multiple Gbps capacities continue to be the topic of extensive analytical and experimental research [10], [9], [11]. Typically, these methods are based on improving the congestion control and buffer management,

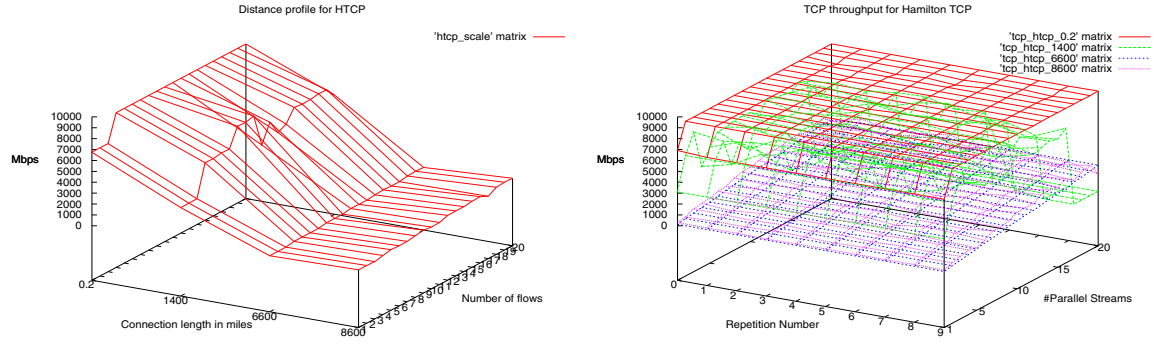
length (miles)	0.2	1400	6600	8600	average
ave throughput 5-15 streams (Gbps)	9.12	6.69	0.76	0.50	4.30
std dev (Mbps)	64.11	70.08	24.96	21.08	33.78
DPM	-	1.74	1.27	1.00	1.34



(a) distance profile (b) stability profile

Fig. 2. Performance of BIC on 0.2, 1400, 6600 and 8600 mile connections.

length (miles)	0.2	1400	6600	8600	average
ave throughput 5-15 streams (Gbps)	9.21	6.71	1.22	1.79	4.72
std dev (Mbps)	12.25	377.42	18.96	128.15	44.58
DPM	-	1.79	1.21	0.87	1.29



(a) distance profile (b) stability profile

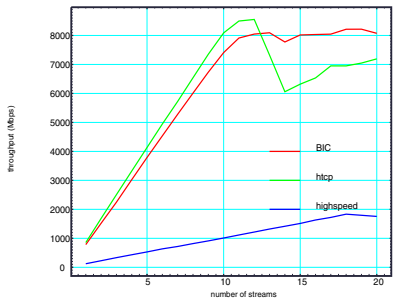
Fig. 3. Performance of HTCP on 0.2, 1400, 6600 and 8600 mile connections.

and a detailed discussion of these methods is beyond the scope of this paper. In addition, there are also several non-TCP methods studied for such connections [18]. The TCP modules used here are readily available in recent Linux kernels.

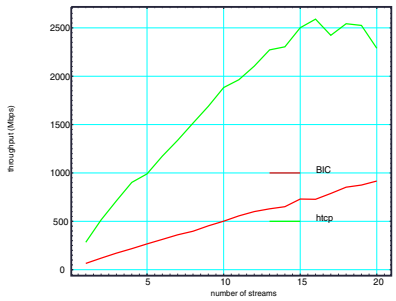
C. Throughput Measurements

The profiles $T_A(d, n)$, for $A \in \{BIC, HTCP\}$ are shown in Figures 2 and 3, respectively, and their comparative performance is illustrated in Figure 4. For single streams at local connections, BIC and HTCP achieved comparable performance of 6.98 and 6.78 Gbps, respectively. For 1400-mile connection, however, HTCP achieved higher throughput of 0.85 Gbps compared to 0.78 Gbps of BIC. For longer connections, both performances are lower with BIC achieving below 65 Mbps, and HTCP achieving 283 Mbps. Clearly, single TCP streams of either type were unable to reach 1Gbps on connections lengths of 1400 miles or longer even with auto-tuning.

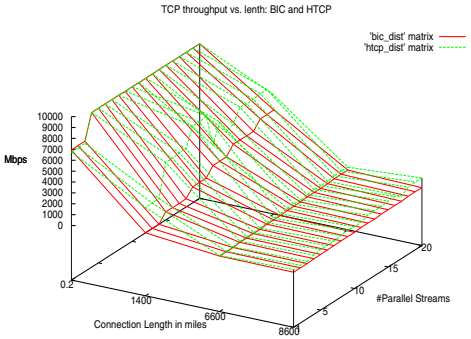
Better throughputs were achieved only using multiple TCP streams, and in both cases, suitable number of multiple streams were needed to achieve the peak throughput. At 1400 miles, BIC and HTCP achieved peaks of 8.2 and 8.5 Gbps using 19 and 12 flows,



(a) BIC, HTCP and HSTCP for 1400 miles



(b) BIC and HTCP for 8600 miles



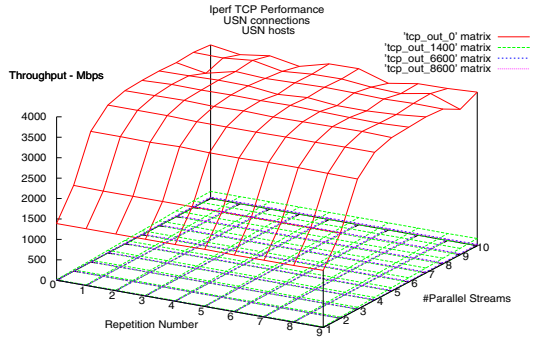
(c) distance profiles of BIC and HTCP

Fig. 4. Comparison of throughputs of BIC and HTCP.

respectively, but when averaged over $n = 15, 16, \ldots 20$, the former achieved higher throughput. Overall, the performance of BIC was higher for 1400 miles when more streams are employed as shown in Figure 4(a), wherein we also included the performance of HSTCP for comparison.

For longer distances, the performance of HTCP was consistently better than that of BIC as shown in Figure 4(b). For 6600 miles, BIC and HTCP achieved peak

miles	0	1400	6600	8600
1 stream (Mbps)	1392	16.32	3.42	2.56
10 streams (Mbps)	3719	176.30	33.50	25.78



(a) stability profile

Fig. 5. Performance profiles for host configuration III

throughputs of 1.39 and 2.38 Gbps using 20 flows each. At 8600 miles, BIC and HTCP achieved peak throughputs of 0.96 and 2.59 Gbps using 19 and 16 flows, respectively. In both cases of BIC and HTCP, the throughput degraded as indicated by DPM values at an overall rate of 1.3Mbps per mile, which is roughly 1.3Gbps loss of throughput for every thousand miles of connection.

Host configurations have a significant effect on the throughputs achieved by the transport methods. To illustrate the extreme case of default TCP throughput, we repeated the measurements using less powerful hosts of Type III in Table I over the same connections with default BIC. The TCP throughput with 10 streams was limited to 3.7 Gbps even for a local connection and was degraded to 26 Mbps for 8600-mile connection as shown in Figure 5. The reasons for this low performance include the slower PCI-X bus together with limited auto-tuning of TCP stack for Linux kernel 2.6.9.

In summary, achieving multiple Gbps throughputs over connections of thousands of miles requires a suitable choice of the host configuration and NIC together with TCP implementation and the corresponding choice for number of streams. It is quite possible to achieve higher TCP throughputs than indicated here by further optimizing various parts, which would in general require a deeper knowledge of the hosts and TCP optimizations.

IV. Wide-Area InfiniBand Solution

The IB Architecture (IBA) [19] defines an open specification for interconnecting compute nodes, I/O nodes and devices in a system-area network. It includes a communication architecture from the switch-based network fabric to transport layer communication interface for inter-processor communication. Processing and I/O nodes are connected as end-nodes to the fabric by HCAs. The collection of nodes and switches is referred to as an IB subnet. IBA specifies transport services and protocols in its communication stack. The OpenFabrics Enterprise Distribution (OFED) supports IB, which is developed and maintained by the OpenFabrics Alliance[20]. OFED includes software packages that support a broad range of environments, including message passing, file system and storage.

A. Extending the Reach of InfiniBand to Wide-Area

A new class of IBoWA devices [14], [15] connect an InfiniBand subnet to the ends of a wide-area connection, thereby extending their IB connectivity to wide-area. However, as IB was originally designed for enterprise-level deployments, its flow control method is limited to the delays of few milliseconds needed for such deployments. To overcome this restriction, the IBoWA devices provide a large amount of buffers at the edge of a wide-area connection, and effectively"terminate" the original IB flow control locally. The network packets from IB subnets are converted by these IBoWA devices to SONET or Ethernet packets, and transmitted over a long-haul connection of thousands of miles. At the same time, these IBoWA devices preserve the native IB flow control dynamics when communicating with the HCAs or switches in the IB networks, thereby achieving the compatibility with native InfiniBand deployments.

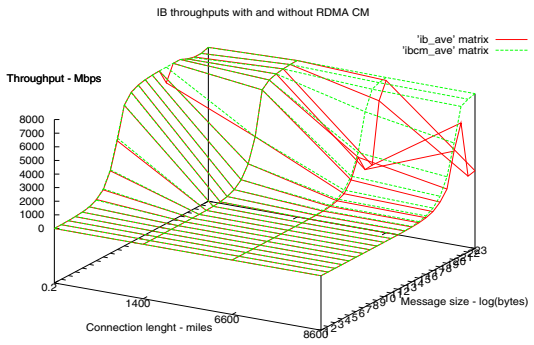
In our tests, the IB solution consists of end hosts with HCAs connected via PCI Express I/O bus, which are connected to IB ports of Longbow or NX5010 devices. These WAN ports of IBoWA devices in turn are connected to wide-area OC192 or 10GigE WAN-PHY or LAN-PHY connection. For SONET, there devices are connected to CDCI switches as shown in Figure 7, and for LAN-PHY they are connected to E300 switches which are in turn connected to CDCI switches as shown in Figure 8. For WAN-PHY, they can be connected to either CDCI or E300 switches; however, the performance in these cases is quite similar to the corresponding SONET or LAN-PHY connections and hence is not discussed here.

B. IB Connection Establishment with RDMA CMA

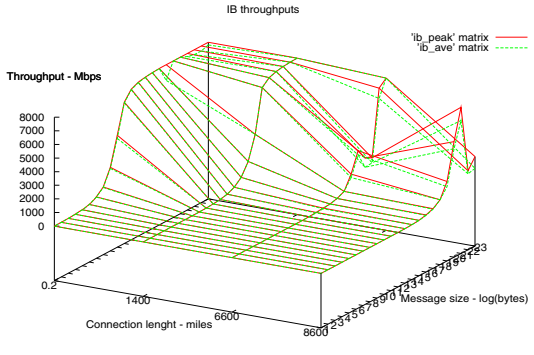
The ib_rdma_bw benchmark provides an option to select two different methods for setting up IB connections. With the default method, two processes establish a socket in advance, exchange their IB connection parameters through the socket, and finish setting up IB connections through its INIT/RTR/RTS phases. If the RDMA CMA method is chosen, the processes set up their connections by invoking the RDMA CM interface. The connection initialization and tear-down are done natively via InfiniBand. In either case, the ib_rdma_bw benchmarks measures the throughput of data communication as seen by 5000 RDMA operations, excluding the time taken for connection establishment. We collected measurements using both methods for connection setup. To our surprise, the performance difference was quite significant at large connection lengths. While the difference was negligible for 1400 miles, the former achieved much higher throughputs as shown in distance profiles in Figure 6(a). Furthermore, the measurements using RDMA CM are more stable, whereas without this option the measurements varied quite significantly leading to non-smooth stability profile as shown in Figure 6. Our initial examination suggests that this could be because the IB connection parameters were chosen differently in two different connection setup methods. The explicit parameters used in the default case were unable to achieve the best performance of InfiniBand on the wide-area network across very long distances.

C. Performance Profiles

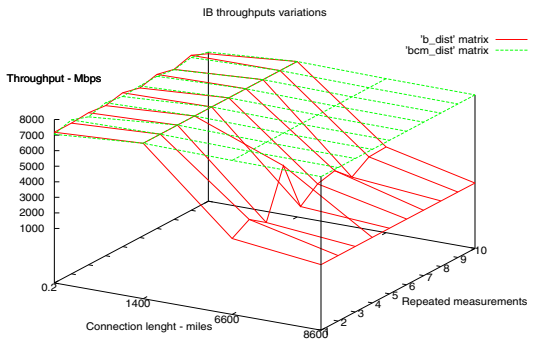
Using the *ib_rdma_bw* benchmark with -c option that utilizes CM to setup the connection, we have generated performance profiles of InfiniBand RDMA by varying the message sizes and connection lengths. Let $T_B(d, s)$, $B \in \{SONET, WANPHY\}$, denote the RDMA throughput measurement collected by ib_rdma_bw tool



(a) average throughput with and without RDMA CM



(b) peak and average throughput without CM



(c) stability profiles with and without CM

Fig. 6. Effects of RDMA CM

over a wide-area connection of length d of type B using a message size of s. Let $\bar{T}_B(d)$ denote the average throughput with 8M message size over 10 repeated measurements on a wide-area connection of type B and length d. As in the case of previous section, we compute the DPM of IB throughput as $D_B(d_i) = \frac{\bar{T}_B(d_0) - \bar{T}_B(d_i)}{d_i - d_0}$.

Three types of performance profiles are generated to characterize the performance.

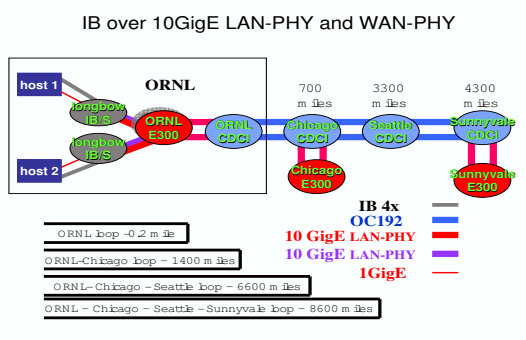
(a) **Distance Scalability:** We generate the IB *distance-profile* $T_B(d, s)$ by measuring the throughput for messages size $s = 2^0, 2^1, \ldots, 2^{23}$ bytes and connection length $d = 0.2, 1400, 6600, 8600$ miles.

(b) **Performance Stability:** We compute the *stability-profile* by repeating 10 times the throughput measurements for fixed message size s =8M, and for $d = 0.2, 1400, 6600, 8600$ miles. We also compute the throughput DPM $D_B(d)$ for these connection lengths.

(c) **Cross-Traffic Profiles:** In IB *cross-traffic profile* we measure $T_B(d, s)$ for fixed d and $s = 2^0, 2^1, \ldots, 2^{23}$ under the presence of cross-traffic on wide-area connection at the rates of 1, 2, 3 and 4 Gbps. We utilize the additional ports on E300 switches to inject external UDP traffic as will be discussed later in this section. We depict the cross-traffic profile as a two-dimensional plot with cross-traffic rate along x-axis and s along y-axis.

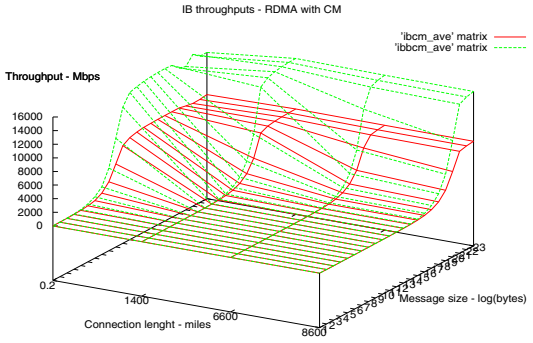
1) SONET Connections: The distance- and stability-profiles for SONET connections are shown in Figure 7. Using SDR 4x HCAs, average throughput of 7.48 Gbps is achieved using Longbow XR devices on the local connection. The throughput only decreased to 7.47 Gbps for 1400 mile connection and to 7.34 Gbps for 8600 mile connection. Furthermore, these throughputs measurements were very stable as indicated by the low standard deviation of measurements in Figure 7. Also, the throughput DPM is at the worst 17Kbps per mile, which is at least 50 times better than the corresponding rate for TCP BIC and HTCP.

2) WAN-PHY Connections: The configuration and performance profiles for 10GigE WAN-PHY connections are shown in Figure 8. The throughput performance in this case is very similar to SONET connection, thereby showing that 10GigE connection is equally viable to support IBoWA devices. This is an important aspect since it has become significantly cheaper to deploy wide-area 10GigE networks compared to SONET networks. In Figure 8(c) we superimposed the distance-profiles computed using the peak and average values of $T_B(n, s)$ – they match quite closely, which is another indication of the stability of measurements.

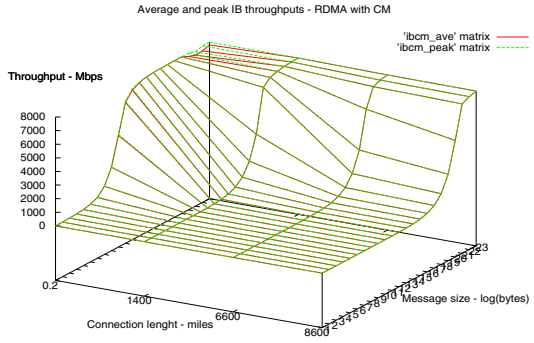


(a) configuration

miles	0.2	1400	6600	8600	ave
ave. throughput (Gbps)	7.50	7.49	7.39	7.36	7.43
std, dev (Mbps)	0.07	0.69	0.08	0.20	0.11
DPM (Mbps/mile)	-	0.018	0.018	0.017	0.017



(b) distance profile of uni- and bi-directional bandwidth



(c) overlapping average and peak bandwidth profiles

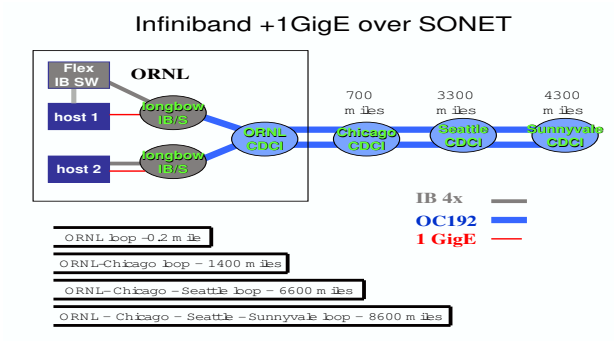
Fig. 8. Performance of IB over 10GigE WAN-PHY on 0.2, 1400, 6600 and 8600 mile connections.

D. Effects of Wide-Area Cross-Traffic

For WAN-PHY connections, we injected UDP cross-traffic using additional USN hosts of type III at rates 1, 2, 3 and 4 Gbps in the configuration shown in Figure 9(a). The cross-traffic profiles are shown in Figure 9(b) for $d = 1400, 8600$. When cross-traffic levels are below 1Gbps, there is no impact on the profile, but throughput was drastically lowered at cross-traffic levels of 2Gbps and higher, and the effect is worse at longer distances. Under the former case, the residual WAN bandwidth after accounting for cross-traffic is above 8Gbps needed for IB DDR 4x, but in the latter case the residual bandwidth is below 7.6Gbps. However, in the latter case, the achieved IB throughput is much lower than the residual bandwidth as shown in Figure 9, for example, below 1Gbps for the 8600-mile connection with 4Gbps cross-traffic.

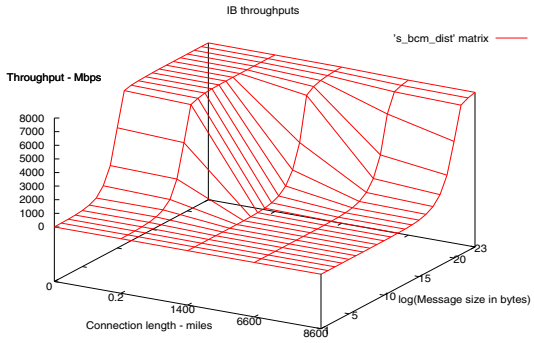
E. 1GigE Support on IBoWA Devices

By utilizing the 1GigE ports on Longbows, we collected throughput measurements while IB traffic was sent at the highest rate for that connection. The IB traffic and 1GigE did not have any interference effect on each other. The TCP and UDP measurements for 1GigE connection that parallels IB connection are shown in Figure 10, which remained the same with or without IB traffic. However, throughputs are higher when only one GigE port is utilized on Longbow XRs. The aggregate throughput when both 1GigE ports are used simultaneously to carry traffic is lower as shown for 1400 mile connection in Figure 10(b).

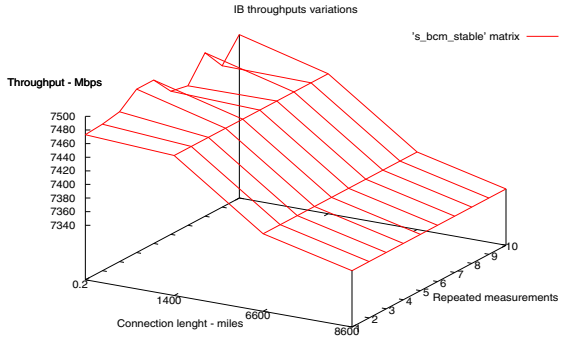


(a) configuration

miles	0.2	1400	6600	8600	ave
ave. thruput (Gbps)	7.48	7.47	7.37	7.34	7.47
std, dev (Mbps)	45.27	0.07	0.09	0.07	11.40
DPM (Mbps/mile)	0	0.012	0.017	0.016	0.015



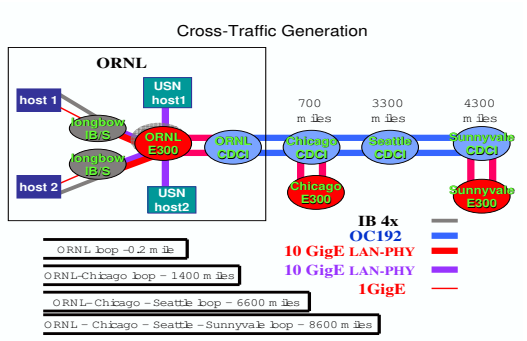
(b) distance profile



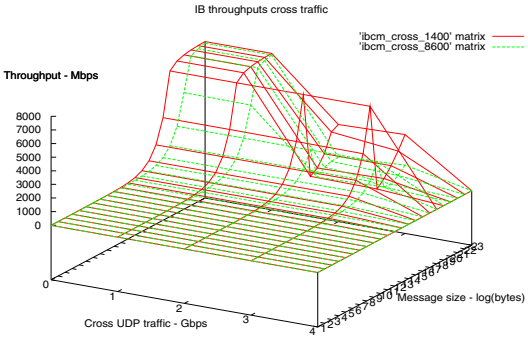
(c) stability profile

Fig. 7. Performance of IB over SONET connections.

miles	1400	6600	8600
0	7.49	7.39	7.36
1G	7.49	7.39	7.36
2G	3.13	1.38	0.74
3G	3.25	1.97	1.02
4G	2.91	1.82	0.96



(a) configuration



(b) cross-traffic profiles

Fig. 9. Cross-traffic effects on IB over 10GigE WAN-PHY on 0.2, 1400, 6600 and 8600 mile connections.

By combining these results with those in previous section, to fully achieve peak IB throughputs, the cross-traffic must be kept at or below 1Gbps whether injected onto WAN connection externally or directly into a single Ethernet port on IBoWA device.

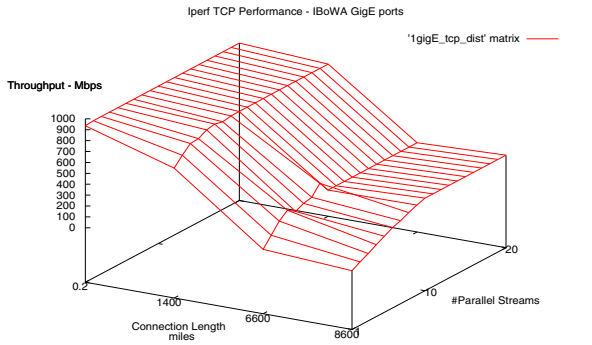
V. Comparison of 10GigE and IBoWA

We now compare the TCP over 10GigE solutions with IBoWA solutions for wide-area data transport in terms of cost, ease of deployment and throughput.

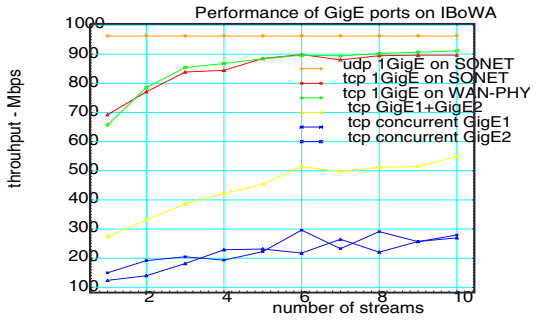
A. Deployment Considerations

In terms of hosts, these two methods have the same order of costs and complexity in that they require suf-

miles	0.2	1400	6600	8600
TCP thruput (Mbps)	944	896	319	350
UDP thruput (Mbps)	962	962	962	962



(a) distance profile



(b) Performance on 1400 mile connection

Fig. 10. GigE traffic supported by IBoWA devices.

ficiently powerful processors and PCI-Express bus to support 10GigE NIC or IB HCA. Also costs of NICs and HCAs themselves are not significantly different.

In terms of edge devices, IB solution requires the special IBoWA devices which represent a significant procurement cost. On the positive side, these devices could be simply dropped in-place, and require very simple configuration changes to switch between OC192 and 10GigE LAN-PHY/LAN-PHY. But they support only point-to-point connections and must be provided at both ends of each connection. In comparison, TCP 10GigE solutions can be implemented on shared wide-area connections very easily. Hosts with 10GigE NICs can be directly plugged into 10GigE edge switches that terminate the wide-area 10Gbps connections. But, the performance tuning of TCP is connection-specific and requires significant expertise and effort.

In terms of wide-area connection, IB solution essentially requires a dedicated wavelength for OC192 or 10GigE WAN-PHY configurations of IBoWA devices. Typically, OC192 connections are cross-connected to wide-area links at full bandwidth and do not support third party cross-traffic. Cross-traffic can be introduced when 10GigE wide-area connections are used by IBoWA devices by trunking other Ethernet traffic along with that due to these devices. But as shown in the previous section, cross-traffic levels above 1Gbps rates can drastically reduce IB throughput. Also, two 1GigE Ethernet ports on a IBoWA device support parallel Ethernet streams, but each with the capacity limited to 1 Gbps. On the other hand, the deployment of 10GigE TCP solutions does not require an exclusive access to entire wavelengths for wide-area connections.

B. Throughput Considerations

For point-to-point data transfers, IB is able to achieve and sustain higher throughputs over longer distances as indicated by the average DPM, given by $\bar{D}_C = \frac{1}{3} \sum_{d=1400,6600,8600} D_C(d)$, of less than 0.02 Mbps per mile, as opposed to 1.3 Mbps per mile for both BIC and HTCP, which represents a scale factor of about 65.

In terms of achievable throughputs , IB performance is very stable as indicated by the standard deviation of 0.09 Mbps on wide-area connections, compared to around 30 Mbps for multiple stream TCP throughputs. For 1400 mile connection, BIC and HTCP with suitable number of parallel streams achieve throughputs above 8Gbps rate, which is above the IB throughput of 7.4Gbps. But when combined with 1GigE traffic supported by IBoWA devices, their aggregate throughput is at the level of 8.2Gbps. On the longer connections, however, data throughputs of IBoWA are about 7.3 Gbps, whereas the HTCP throughputs were limited to 2.5 Gbps.

IB solution does not gracefully degrade in the presence of cross-traffic above few Gbps. Even short duration cross-traffic levels of 4 Gbps significantly degrade the IB performance. TCP on the other hand has the built-in mechanisms to adapt to decreases in available bandwidth due to cross-traffic.

VI. Conclusions

There are two seemingly different approaches for achieving wide-area high-performance data transfers over connections of thousands of miles based on: (a) 10GigE technologies supported by TCP, and (b) IB technologies supported by their wide-area extensions. We compared the performance profiles of both solutions over various 10Gbps connections of lengths up to 8600 miles using off-the-shelf systems. Such profiling is a first step towards assessing the performance of high-performance applications in that their throughputs will be upper-bounded by the profiled measurements. Our results illustrate the complexity of deploying these technologies and the need for optimizations in realizing and sustaining such end-to-end throughputs. The comparative performance between these two approaches leads to multi-faceted trade-offs. For data transport using a single large flow and the rest of flows aggregated to 1 Gbps rate, IB solution is better suited. On the other hand, for multiple competing flows on wide-area connection, 10GigE solution is better, particularly at shorter distances.

Our performance profiling results are primarily targeted towards data transport and do not necessarily reflect the performance of more complex applications, for example, MPI-based computation distributed across two remote supercomputers or a high-performance file system mounted across a wide-area connection. In general, latency-sensitive applications may have to be suitably enhanced to account for the larger RTT of wide-area connections. For example, additional adaptations and tuning would be required to achieve high file transfer rates for the configurations presented in this paper, and it would be of interest to examine the measurements from the corresponding file system benchmarks. It would also be of interest to select benchmark applications involving real-time tasks such as instrument control, computational steering and visualization, and generate their distance- and stability-profiles. Finally, it would be of future interest to connect supercomputers, storage and files systems over multiple IBoWA connections and generate the performance profiles for various data transfer applications.

VII. Acknowledgments

Authors are grateful to Pete Wyckoff for his comments that greatly improved the discussion and presentation of the results in this paper. This research is sponsored by the High Performance Networking Program of the Office of Science, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC, and by Department of Defense.

References

[1] "Internet2," http://www.internet2.edu.

[2] "Energy Sciences Network," http://www.es.net.

[3] N. S. V. Rao, W. R. Wing, , S. M. Carter, and Q. Wu, "Ultrascience net: Network testbed for large-scale science applications," *IEEE Communications Magazine*, 2005, in press, expanded version available at www.csm.ornl.gov/ultranet.

[4] End-To-End Provisioned Optical Network Testbed for Large-Scale eScience Application, http://www.ece.virginia.edu/ mv/html-files/ein-home.html.

[5] L. Xu, K. Harfoush, and I. Rhee, "Binary increase congestion control (bic) for fast long-distance networks," in *INFOCOM*, 2004.

[6] I. Rhee and L. Xu, "Cubic: A new tcp-friendly high-speed tcp variant," in *Proceedings of the Third International Workshop on Protocols for Fast Long-Distance Networks*, 2005.

[7] S. Floyd, "Highspeed TCP for large congestion windows," 2003, internet draft, February 2003.

[8] R. Shorten and D. Leith, "H-TCP: TCP for high-speed and long-distance networks," in *Proceedings of the Third International Workshop on Protocols for Fast Long-Distance Networks*, 2004.

[9] T. Kelly, "Scalable tcp: Improving performance in high speed wide area networks," *Computer Communication Review*, vol. 33, no. 2, pp. 83–91, 2003.

[10] M. Hassan and R. Jain, *High Performance TCP/IP Networking: Concepts, Issues, and Solutions*. Prentice Hall, 2004.

[11] D. Kitabi, M. Handley, and C. Rohrs, "Congestion control for high bandwidth-delay product networks," in *Proc. of SIGCOMM*, 2002.

[12] "bbcp," http://http://www.slac.stanford.edu/ abh/bbcp/.

[13] "Gt 4.0 gridftp," http://www.globus.org.

[14] Obsidian Resarch Corporation, http://www.obsidianresearch.com/.

[15] Network Equipment Technologies, http://www.net.com.

[16] N. S. V. Rao, W. R. W. S. E. Hicks, S. W. Poole, F. A. Denap, S. M. Carter, and Q. Wu, "Ultrascience net: High-performance network research test-bed," in *International Symposium on on Computer and Sensor Network Systems*, 2008.

[17] W. Yu, N. S. V. Rao, and J. S. Vetter, "Experimental analysis of infiniband transport services on WAN," in *International Conference on Networking, Architecture and Storage*, 2008.

[18] E. He, P. V. Primet, and M. Welzl, "A survey of tranport protocols other than "standard" TCP," global Grid Form Report GFD-I.055.

[19] Infiniband Trade Association, http://www.infinibandta.org.

[20] OpenFabrics Alliance, http://www.openfabrics.org.